



因果関係を考慮した連鎖パターンマイニング手法のパラメータの設定

メタデータ	言語: Japanese 出版者: 計測自動制御学会 公開日: 2016-02-17 キーワード (Ja): シーケンシャルパターンマイニング, 連鎖パターンマイニング, 因果関係, データマイニング キーワード (En): 作成者: 大久保, 勇輔, 李, セロン, 岡田, 吉史 メールアドレス: 所属:
URL	http://hdl.handle.net/10258/3861

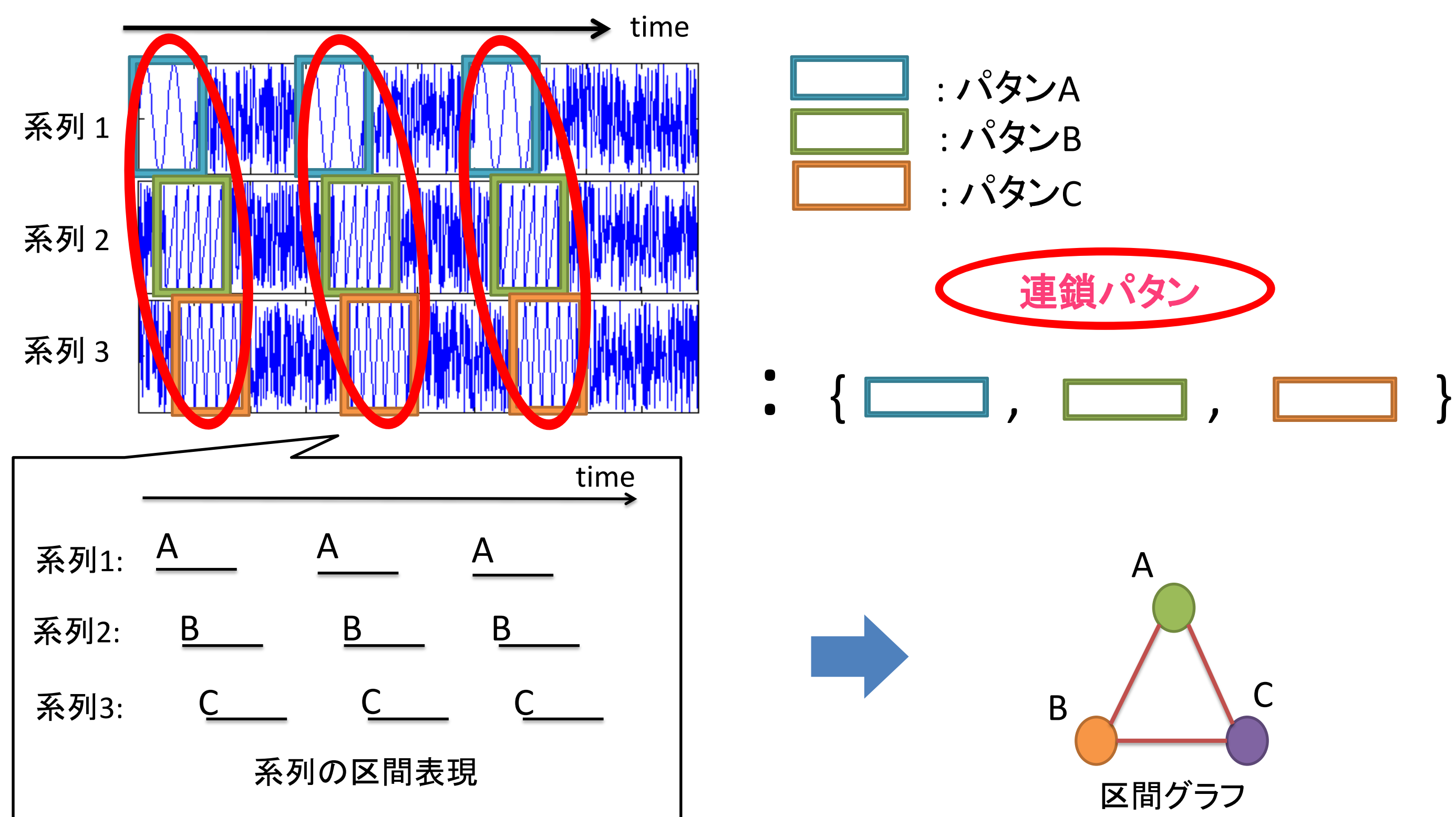
因果関係を考慮した連鎖パターンマイニング手法のパラメータの設定

著者	大久保 勇輔, 李 セロン, 岡田 吉史
雑誌名	計測自動制御学会システム・情報部門学術講演会講演論文集
巻	2015
発行年	2015-11-18
URL	http://hdl.handle.net/10258/3861

背景

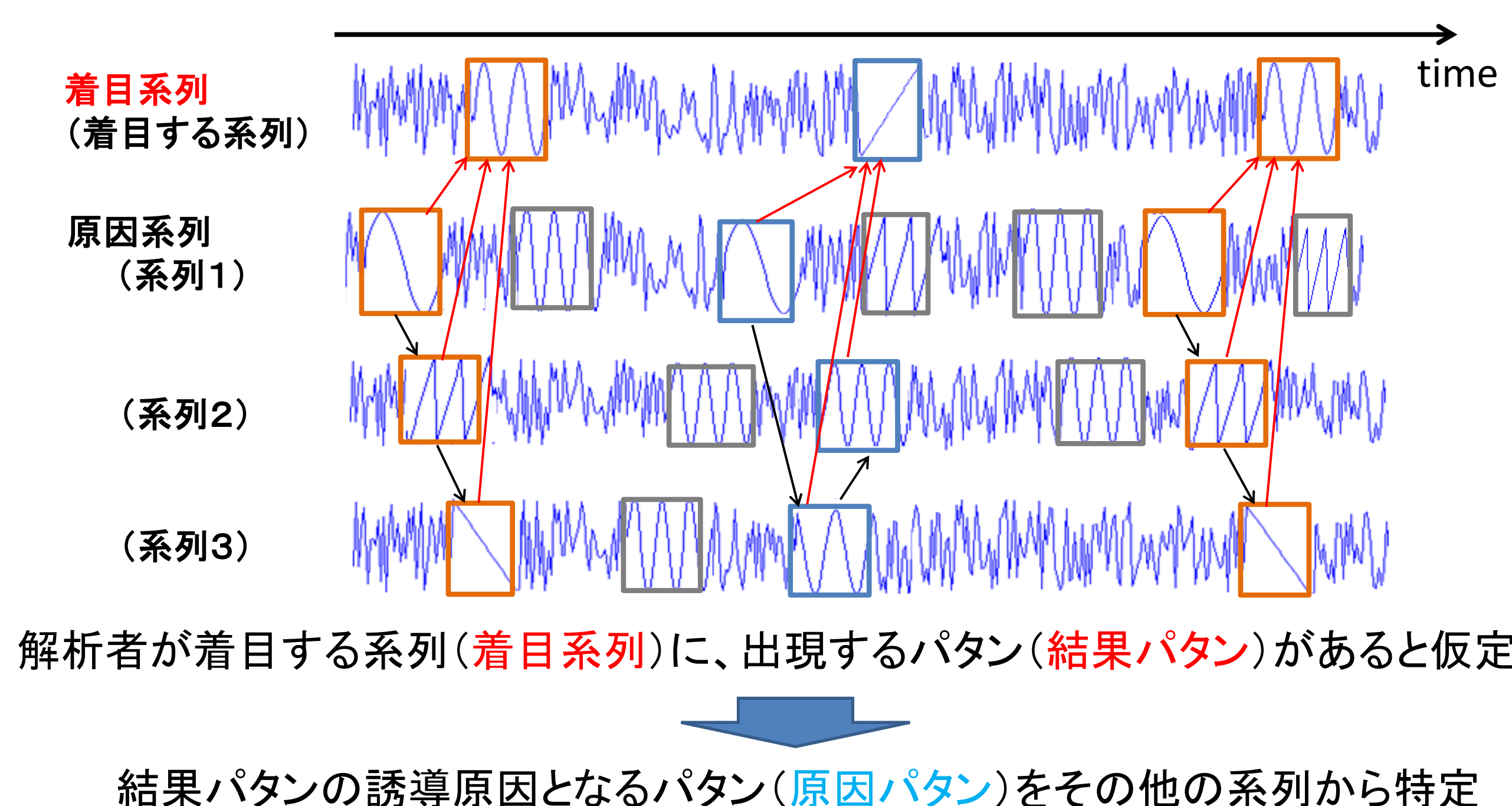
▽連鎖パターンマイニング手法

複数の系列間にまたがって、同時刻帯に繰り返し現れる頻出パターン(連鎖パターン)を抽出する手法 (2012. miura)



本研究の目的

以前の連鎖パターンマイニング手法を拡張して、
複数の系列データ(脳波、脈波など)から、興味の事象の発生原因を見つけ出す



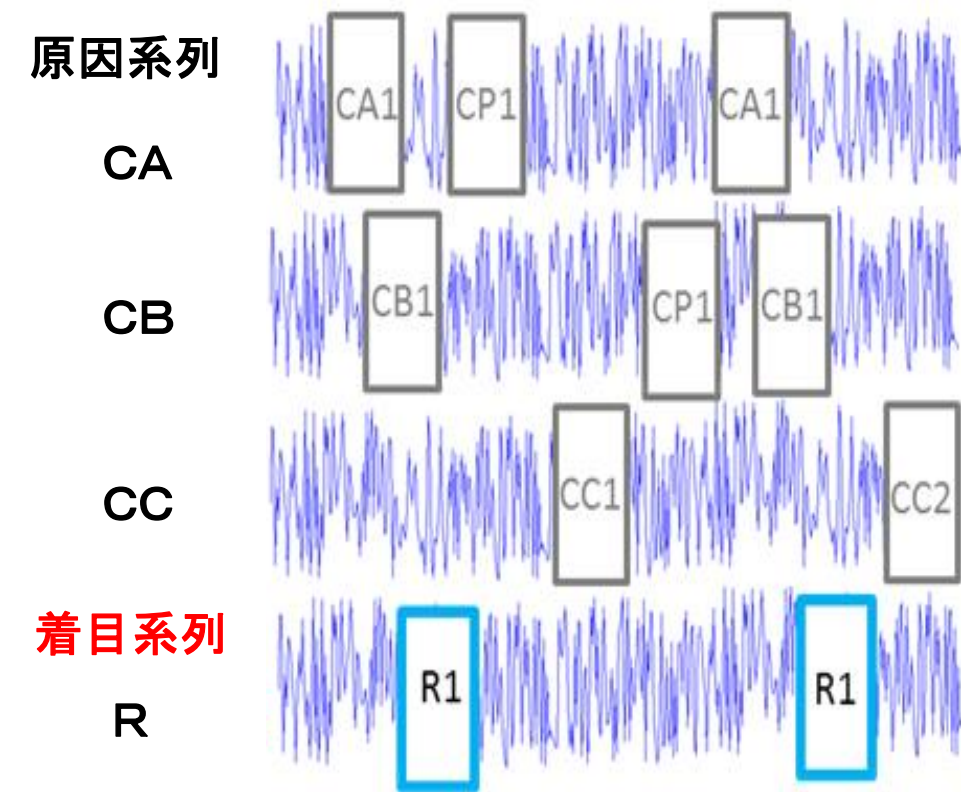
本研究の応用例

- 複数のバイタルデータ(脳波、心電図)から、病的特徴を抽出
- バイタルデータ、音声データから感性情報を解析

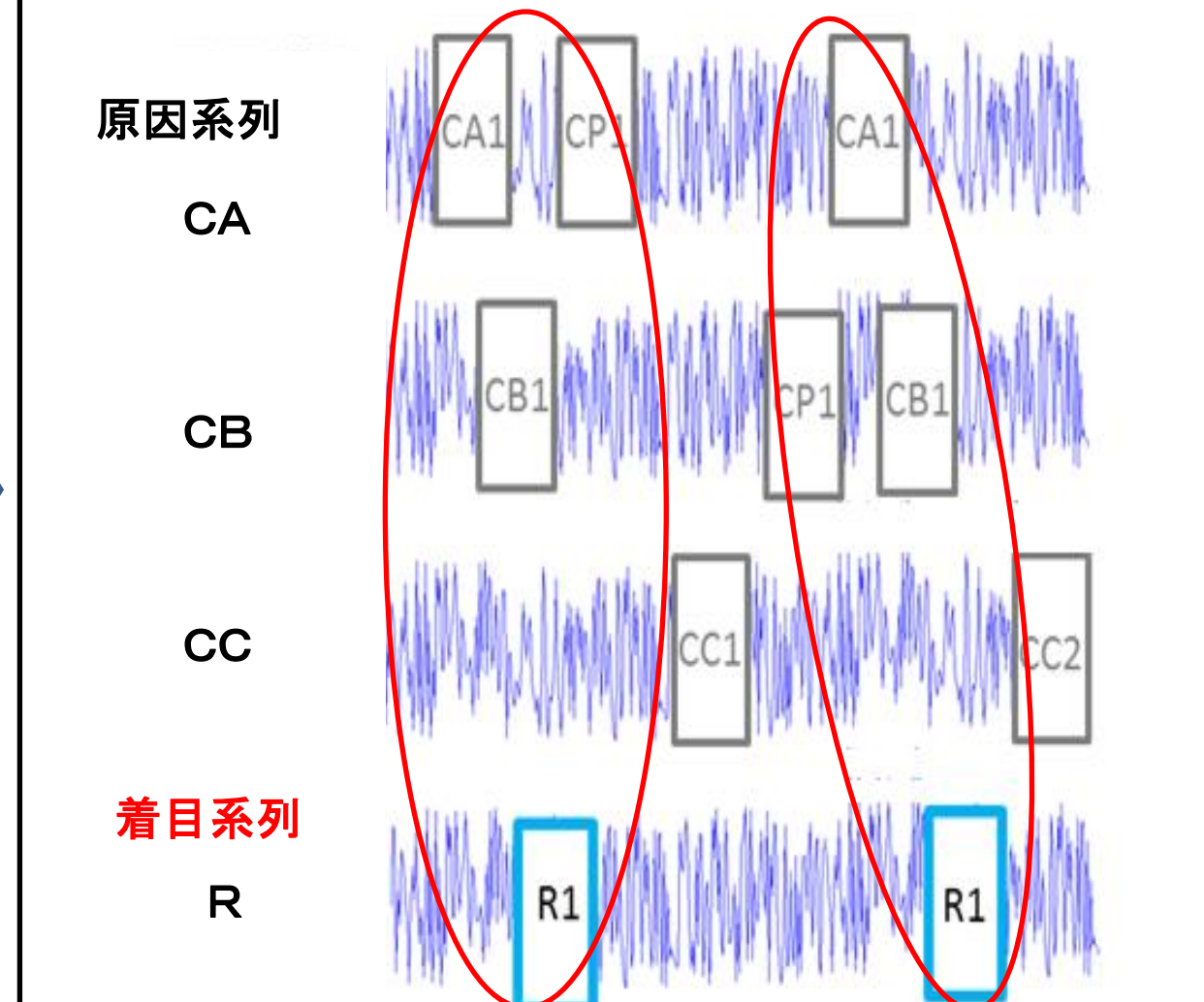
手法

□手法の流れ

Input: 系列データ



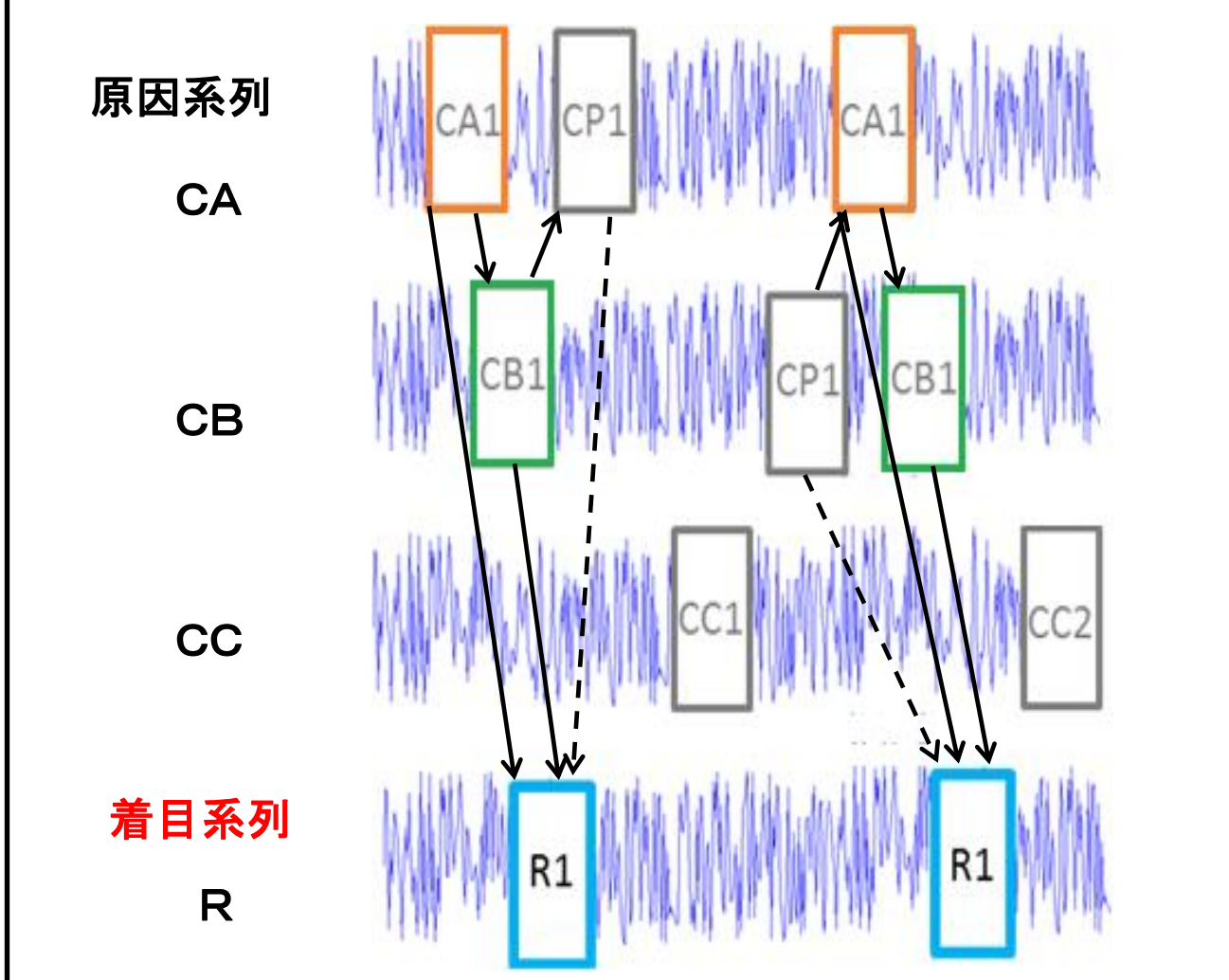
Step1: 連鎖パターンを抽出



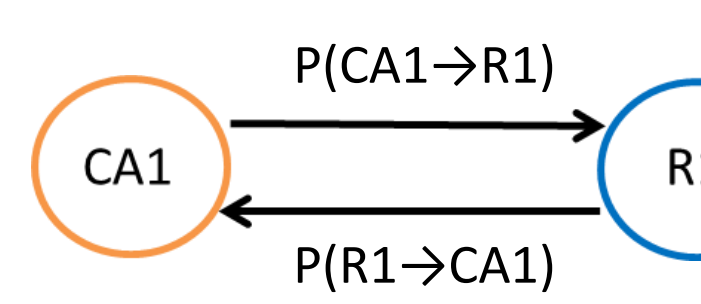
- ・前手法を用いて連鎖パターンマイニングを行う
- 同時刻帯に重なって出現する頻出パターンを連鎖パターンとして抽出

連鎖パターン: {CA1, CB1, R1, CP1}
{CP1, CA1, CB1, R1}

Step2: 因果の確率を計算



原因→結果 (順方向)

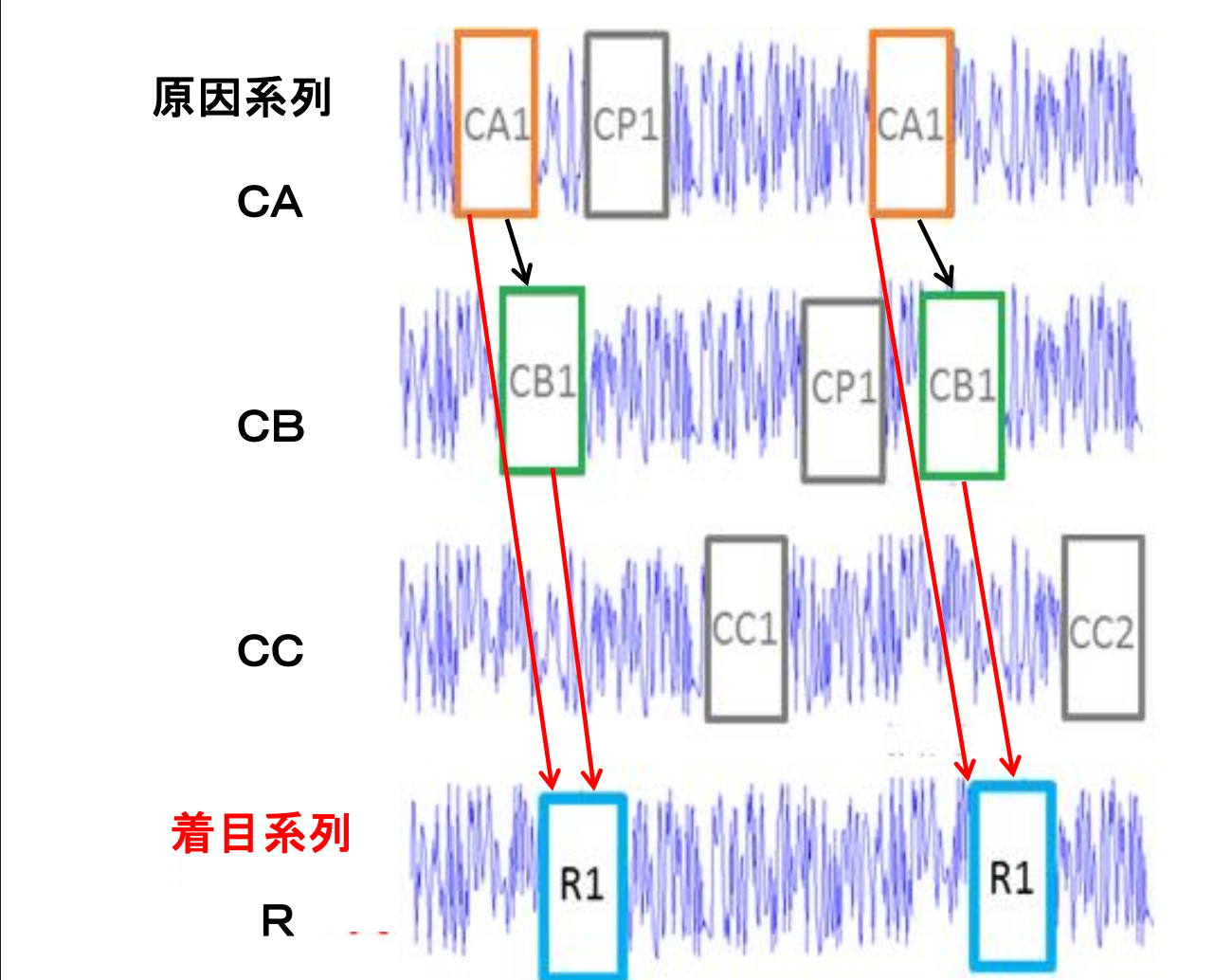


結果→原因 (逆方向)



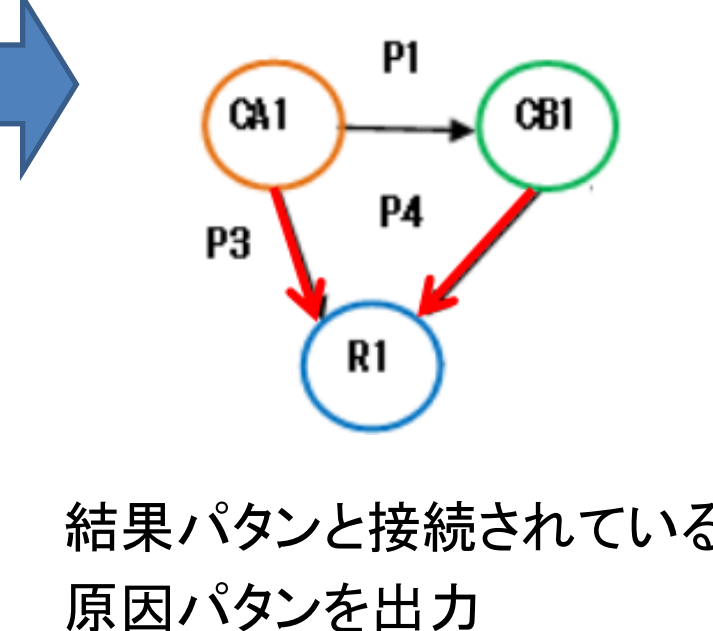
- ・CA1が発生した条件下でR1が発生する確率
T1: 順方向の確率の閾値
- ・R1が発生した条件下でCA1が発生する確率
T2: 逆方向の確率の閾値

Step3: 因果関係のないエッジを削除



Output:

原因パターン: CA1, CB1
結果パターン: R1



評価実験

□人工データセットを用いて抽出精度を評価

- ・一様乱数による系列データに連鎖パターン(正解パターン)を2種類とノイズとなる頻出パターンを埋め込んだ

□評価指標

$$\text{適合率(Precision)} = \frac{CDP}{DDP} \quad \text{再現率(Recall)} = \frac{CDP}{EDP}$$

$$F \text{ 値} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

CDP: 正しく抽出された正解パターンのデータ点数

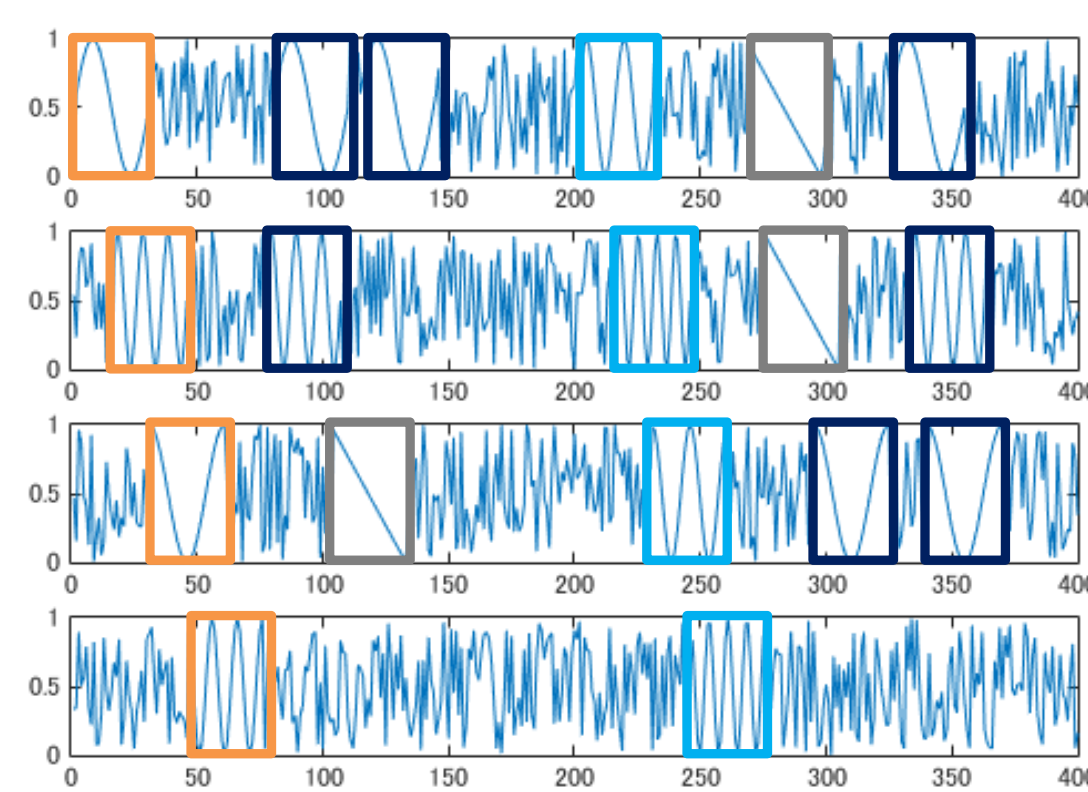
DDP: 本手法により連鎖パターンとして抽出された部分のデータ点数

EDP: 正解パターンのデータ点の総数

□データセットの種類

種類	データセットの内容
dataset1	パターンの長さを変更 (dataset1.1, dataset1.2, dataset1.3)
dataset2	正解パターンの出現間隔と各系列の長さを変更 (dataset2.1, dataset2.2, dataset2.3)
dataset3	正解パターンの出現頻度を変更 (dataset3.1, dataset3.2, dataset3.3)
dataset4	ノイズパターンの出現頻度を変更 (dataset4.1, dataset4.2, dataset4.3)
dataset5	ノイズパターンに正解パターンの一部(ダミーパターン)を挿入し、ダミーパターンの出現頻度を変更 (dataset5.1, dataset5.2, dataset5.3)

□人工データの一部

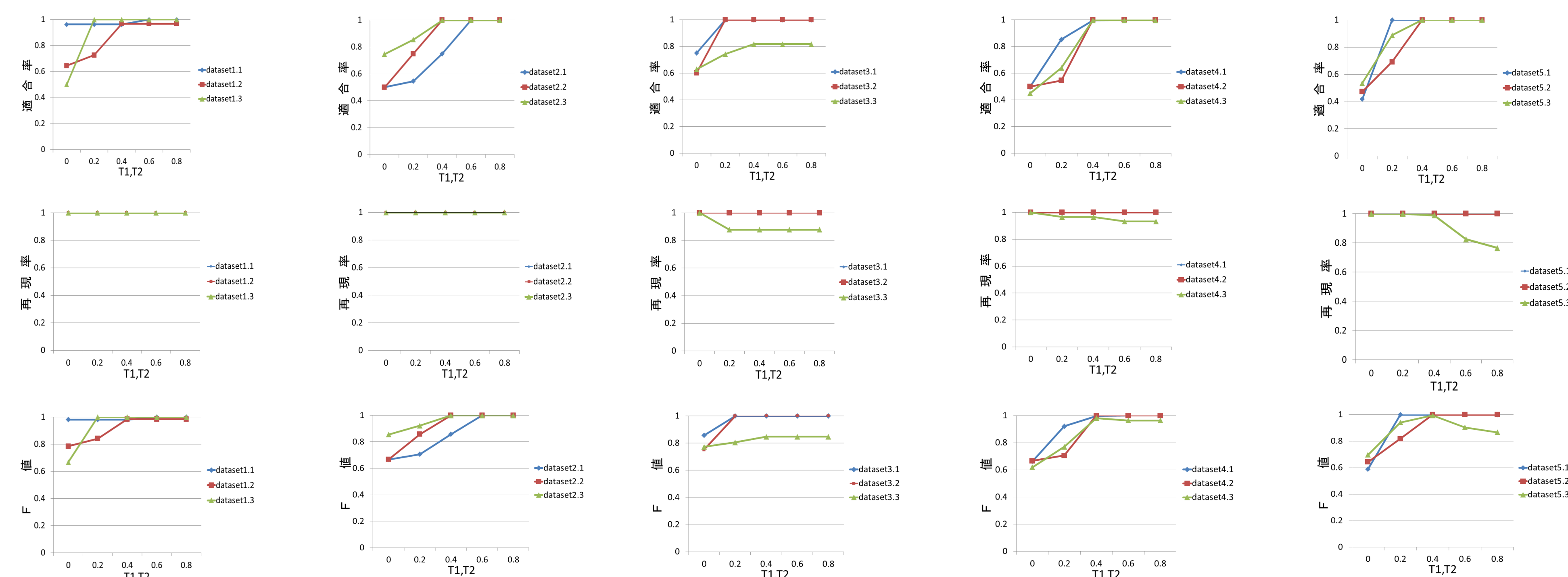


正解パターン

ノイズパターン

ダミーパターン

結果と考察



○ dataset1~3, 4.1, 4.2, 5.1, 5.2

T1, T2の値が大きいくほどF値は高くなる

⇒ 結果パターンとは無関係な原因パターンを連鎖パターンから削除することにより、因果関係のあるパターンで構成される連鎖パターンのみ正しく抽出している

○ dataset4.3, 5.3 (ノイズパターン、ダミーパターンの出現頻度が高い場合)

T1, T2=0.4のとき、F値がピークとなりそれより大きなT1, T2では抽出精度が低下
⇒ ノイズパターン、ダミーパターンの影響を受けて正解パターンを正しく検出できなくなる

○ dataset2.1, 3.3 (正解パターンの出現間隔が短い or 出現頻度が高い場合)

T1, T2が低く設定されたとき、他のデータセットより抽出精度が低下
⇒ パターンが密に埋め込まれ、ノイズパターンが混入しやすくなるため

まとめ

○パターンが密に埋め込まれているデータセットでは、T1, T2の値を大きく設定することにより、抽出精度が高くなる

○ダミーパターンが含まれたデータセットは、知見が得られなかった

□今後の課題

- 系列データの性質に応じて、閾値を設定する方法の開発
- 実データへの適用実験を行い、よりノイズに頑健な手法へと拡張を目指す